Chemometrics and Intelligent Laboratory Systems 149 (2015) 66–72

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Assumption free modeling and monitoring of batch processes☆

Frank Westad [1], Lars Gidskehaug [1], Brad Swarbrick [1], Geir Rune Flåten [1]

CAMO Software AS, Postboks 1662 Vika, 0120 Oslo, Norway

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form 31 July 2015
Accepted 27 August 2015
Available online 5 September 2015

Keywords:
Batch processing
Batch modeling
Multivariate modeling
Batch analysis
Process analysis
Batch monitoring
Quality monitoring
Relative time modeling

ABSTRACT

Modeling strategies currently in use for the monitoring of batch processes where multivariate data are available have some limitations, particularly for batches where the true starting or end point are not the same on an absolute time scale, or the batch progression varies among batches. In this paper, a method capturing these differences and allowing modeling and monitoring of batches in relative time is proposed. Using scores from principal component analysis (PCA) models as a feature space the new methodology is better able to handle the challenges usually experienced in batch analysis. The feasibility of the relative time approach is demonstrated using data from a chemical synthesis and a pharmaceutical drying process.

1. Introduction

Batch processes are widely used in many industries. Typically, raw materials are combined in a suitable batch vessel before chemical, physical, or biological transformation takes place, resulting in an end product. In many cases, the control of the batch process is recipe driven, and the operations are not adjusted to accommodate raw material variation, changes in uncontrollable factors, and other changing circumstances. The best possible end product quality is achieved by adapting batch operations according to any detectable changes during processing, thus providing a control mechanism to drive a product toward its desired state. Optimal run settings and the ability to control them within a design space leads to reduced rework and rejects and improved end product quality, which has the major benefit of saving industry money and resources and more importantly, increased consumer trust in the product name.

Several solutions already exist for batch monitoring and control [1–3]. The starting point for all of them is a data set $\underline{\mathbf{X}}$ ($m \times n \times k$) describing normal operation conditions (NOC) of known batches of high quality, where n variables are measured at m sampling times for each of the k batches. The data are structured in two-dimensional matrices along time points $m \times (n^*k)$ or variables $(m^*k) \times n$ unless a direct three-way modeling approach such as PARAFAC [4] is applied. A model describing NOC is developed and new batches can be projected on the model to provide real-time quality information.

The existing batch modeling approaches assume equal lengths of batches, i.e. the batch is expected to start at the same chemical or biological time t_0 and has the same number of time points for all batches. This leads to problems during model building if the NOC data set is uneven, and ultimately during monitoring if new batches do not meet these criteria. Numerous approaches to handle uneven batch lengths exist, including replacing time with a maturity index [2], dynamic time warping (DTW) [5], time linear expanding/compressing [6] etc. Complications can occur in all of these methods if the first measurement does not coincide with the true t_0, i.e. the new batch(es) does not start at the same chemical/biological state. The PARAFAC approach models the data as a true three-way model which has a possible advantage that the time is modeled as a separate dimension and not connected to either samples or variables as in the unfolding case. However, the challenges with unequal batch length and chemical time still need to be addressed. Also, the monitoring phase requires dynamic recalculating of models up to the current point of time [4].

Another challenge experienced in batch monitoring relates to batches where there are phase changes, i.e. the underlying dynamics in the batch changes. This can occur in processes such as multiple stage chemical reactions where a reactant is added during operation to initiate a subsequent reaction. An equivalent situation may occur in physical processes where the underlying mechanism changes, such as in a pharmaceutical wet mass drying process where there is a state change in the material when all free water is evaporated and bound

☆ Selected papers presented at the 3rd European Conference on Process Analytics and Control Technology, 6–9 May 2014, Barcelona, Spain.

E-mail address: grf@camo.com (G.R. Flåten).
[1] Tel.: +47 22 39 63 00; fax +47 22 39 63 22.

http://dx.doi.org/10.1016/j.chemolab.2015.08.022
0169-7439/

water of crystallization starts evaporating. There are approaches for handling such challenges, e.g. multiple local models [5], which can work fine if the phase change occurs at fixed times or a trigger can be used to switch models.

When monitoring batches with phase transitions, varying residence times or reaction rates within the phases may occur. This is often seen in biological batch processes, where bacteria metabolize reactants into products. The modeling solutions applied to the varying residence time are often the same as those used for phase changes, such as multiple local models. However, the uneven progression among batches means that visualizing the trajectory for a batch compared to NOC is not very meaningful, and the presentation is not representative of the underlying batch dynamics. Plotting scores from a multivariate model versus a time axis imposes a one-dimensional assumption on the batch modeling which may not necessarily be correct in relation to the chemical/biological state of the material.

During monitoring of new batches, another challenge experienced by models for data unfolded along the batch dimension is the handling of missing values [7]. At any time during a batch, only the current and previous measurements are available. Measurements at future time points of the batch are missing, and any missing values cannot be directly compared with the established model for NOC. The typical solution for this is to use what is known as lagged multivariate models (LMV), which in effect is a number of models for the different time points [8]. However, this seems to be a work-around solution, although the computational time is often no issue with today's computers for such a complex modeling strategy.

All the above-mentioned approaches effectively try to capture the batch trajectories as described in the multivariate space, making the assumption that time is an attribute of the trajectory. However, [absolute] time is not a necessary attribute of the trajectory [9]. Nevertheless, any process is of course operated along the time dimension. A good modeling approach requires a flexible but true synchronization between trajectory [process state] and time [from a chemical/biological state]. This is a similar approach to the desired state as defined in the quality by design (QbD) paradigm, i.e. processes should not be run for a particular defined time interval, they should be run until the product reaches its desired state (provided the time to reach the end point is not excessive). Thus, a time-independent approach to batch modeling also meets the criteria of the QbD approach.

In this paper, an improved batch modeling approach accommodating uneven batch lengths, unknown true t_0, phase changes, and uneven residence times is proposed. This is achieved by a true multivariate, feature-based approach that does not make any assumptions about the synchronization and duration of batches. Instead, the so-called relative time is estimated by the method itself. Relative time is here used in a broad sense for any transient process including non-linear behavior, and it is often found to correspond with the underlying chemical, biological, or physical changes during the process. The presented method is analogous to existing methods [1–4] in that a calibration set of batches with relevant measurements are used to establish a model representative of NOC. The $m \times n \times k$ calibration set $\underline{\mathbf{X}}$ has data for k batches with m samples in each batch for n variables (where the value of m is not necessarily constant across batches).

The calibration set $\underline{\mathbf{X}}$ is unfolded so that all the observations for each batch are represented as rows and the unfolded data matrix thus has the dimension $(k^*m) \times n$. A principal component analysis (PCA) model for the unfolded data set is calculated and validated. In the resulting score space, a grid is optimized to capture the features of the batch trajectories. A new PCA model is developed based on the samples within the feature grid only, and true trajectories, relative time, and relevant deviations are calculated. It is important at this stage to ensure that the process signature of the trajectories is consistent between batches in order to develop a robust and representative grid. This is highly important and it has been noted as a major flaw in a straight mathematical fitting of data, i.e. if the process trajectories are not visibly overlaying to a high degree, then the process itself is not consistent, therefore a batch model is not representative of tight manufacturing controls.

Monitoring of new batches is done by projection onto the feature grid, which is now based on sound scientific modeling. The feature grid describes NOCs and since it is feature based, any time shifts, phase changes, or rate variations are taken into account by the model. Deviations from the NOC are identified using established diagnostics tools, including a modified Hotelling's T^2 chart that is dynamic over the process trajectory and also a dynamic limit for the F-residual distance to model.

The theory section outlines the relative time approach and the proposed diagnostics plots are discussed in further detail. In the subsequent section, two short examples where the method is applied are shown. The discussion section suggests where the proposed method will be suitable as well as providing a review of the underlying assumptions.

2. Theory

Data collected over time for a number of batches and variables can be represented as a three-way data structure, $\underline{\mathbf{X}}$, of dimensions $(m \times n \times k)$. The indices correspond to m time points for which the n variables are measured for k batches. However, due to various batch lengths and other uncontrollable variations, the sample number within the individual batches may not pertain to the same relative state of the process in terms of the underlying chemistry between batches. Therefore, a more appropriate term for the "time" dimension is "sample number." An example of this is the case where the sampling rate is varied between the batches and thus the sample number does not reflect the batches' development over time, even if the relative states were the same. The situation where the sample number does not reflect the same state is more a rule than an exception for chemical and biological systems.

Some scenarios in addition to the above-mentioned sampling-rate situation are:

a. The batches do not start in the same state, e.g. the material processed has various moisture content at the start of the process.
b. The batches' final sampling points do not reflect the same state.
c. The batch progression is not the same and is in most processes non-linear, i.e. the batch trajectory does not evolve in equidistance steps in relative time.

The solution for handling all the above cases without extensively distorting the relative time by preprocessing is to model the batch trajectory from a common start to a common end state in chemical/biological time.

The approach proposed here comprises the following steps as presented below in pseudo-code:

1. Unfold the $(m \times n \times k)$ data matrix $\underline{\mathbf{X}}$ to the $(k^*m) \times n$ data matrix.
2. Preprocess, center, and scale the data as required.
3. Perform PCA for all batches in one model while cross-validating across batch $\mathbf{X} = \mathbf{t}_a\mathbf{p}_a^t + \mathbf{E}$, a $= 1..A_{optimal}$.
4. Use a grid-search algorithm in the multivariate (score) space to find the common relative start and end points in so-called "grid elements." Various grid resolutions are applied to model the trajectory in the best way. The current criterion is to select the resolution that gives the most grid elements. An alternative would be the highest percentage of grids from the total number of grids. In principle, the grid search can be extended to 3D ("grid cubes") and higher dimensions.
5. Calculate the mean for all samples and the means for individual batches for these grid elements. If not all samples are found inside the grids, recalculate the PCA model.
6. Interpolate to a common batch trajectory for the overall mean with the desired resolution, e.g. in percent of relative time. Linear or

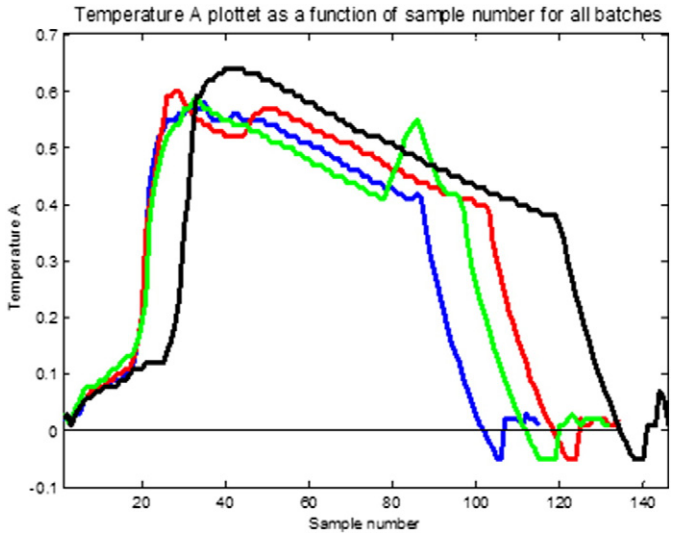

Fig. 1. Variable (Temperature A) plotted vs. sample number.

spline interpolation may be applied depending on the nature of the process.

7. Project the mean for the individual batches onto this trajectory and estimate the relative time, distance within the model space, as well as the residual distance. The projection is orthogonal to the line between two adjacent points on the trajectory and gives the relative time as the ratio of actual point and the total number of points on the trajectory. For simplicity, it is scaled to 0–100.
8. From these projected distances estimate the standard deviation around the common trajectory. The actual points to be plotted as lower or upper limits are always following the direction of the trajectory. This is to ensure that the limits are not crossing when plotting connection lines as shown in Figs. 5, 7, and 10 below.
9. The residual distance for each grid element is also estimated and is thereafter interpolated to yield a residual distance limit for each relative time. The residuals for individual objects are actually the same as the normal residuals in the PCA model but are plotted for the relative time and its actual limit.
10. The individual samples' scores are projected onto the trajectory for estimation of the relative time, distance to the trajectory, and distance to the model.

An alternative would be to perform PCA on the individual batches with individual centering and scaling. However, due to rotational ambiguity, this would require reflection and/or rotation of the models to a common basis before estimating the average trajectory. Also, the objective is mainly to model the differences between the batches and therefore centering and scaling are performed column-wise for the variables. If the process changes character for the whole duration, it might be split into phases to account for various correlation structures. See also section Discussion.

The model is a representation of the dynamics of the batch, in terms of the mean trajectory, so the individual batch mean values are projected onto the nearest point on the trajectory. The average of these distances is defined as the "distance to the model trajectory" and can be displayed at various significance levels. The orthogonal distance to the trajectory model for the objects for a new batch can be compared to the distance limit for monitoring purposes. Given that $t_{new} = x_{new}P$, the distance can be expressed as

$$D_{Trajectory} = \sqrt{\sum_{a=1}^{Aopt}\left(t_{new,a} - t_{new,a\perp}t_{Trajectory,a}\right)^2} \quad (1)$$

Where

$D_{Trajectory}$ is the orthogonal distance from the new score to its projected position on the trajectory
t_{new} is the new score
$t_{new} \perp t_{Trajectory}$ is the projected position on the trajectory

The monitoring phase involves the following steps:

1. Preprocess, center, and scale the new observations
2. Estimate the new scores as $t_{new,a} = \boldsymbol{x}_{new}\boldsymbol{p}_a$ for components 1:$A_{optimal}$
3. Project these scores onto the trajectory for estimation of the relative time, distance to the trajectory, and distance to the model.

Note that this approach estimates batch progression or maturity directly in terms of the relative time.

3. Experimental and results

Two examples will be shown to illustrate the challenges presented with time-dependent processes, both in the modeling phase as well as for monitoring. The examples concern two unit operations from the chemical and pharmaceutical production environments, one with a small number of process variables and one with instrumental data from NIR spectroscopy.

3.1. Example 1: A chemical reaction

The data were taken from a chemical synthesis which, due to a confidentiality agreement, could not be described in detail. The variables were temperature measured at two positions (A and B) and pressure. This example, although only with three variables, illustrates very well the challenges in batch modeling.

For simplicity, the samples were pre-screened to describe the reaction itself and not the initial phase in the synthesis. Nevertheless, although the samples have the same relative starting point, the development over time is quite different. Fig. 1 shows the Temperature A variable as a function of the sample number. The immediate interpretation is that the batches are different, but the 2D score plot from PCA calculated for the three variables (Fig. 2) reveals that the batches follow the same trajectory although not with the same relative rate. Batch 4 has more observations as seen in Fig. 1, which gives the effect of a very condensed cloud of points on the left-hand side in Fig. 2. Batch 3 reveals that the pressure is increasing and thereafter decreasing, yielding the reversed evolution around score value 1 (PC1). The corresponding correlation loading plot [10] revealed that all variables

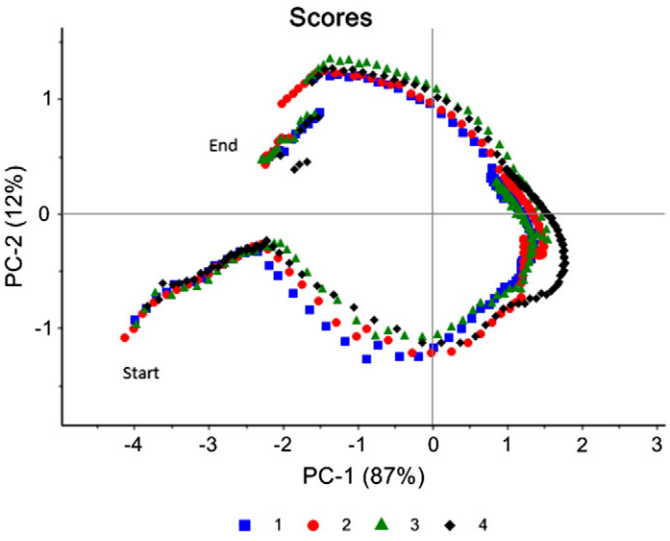

Fig. 2. Score plot of PC1 vs. PC2 with batches shown as different symbols.

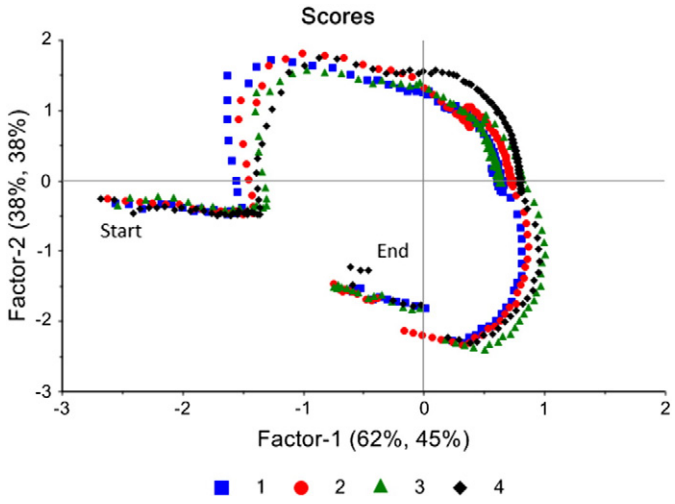

Fig. 3. Score plot from the PLS model for maturity with samples representing common start and end points.

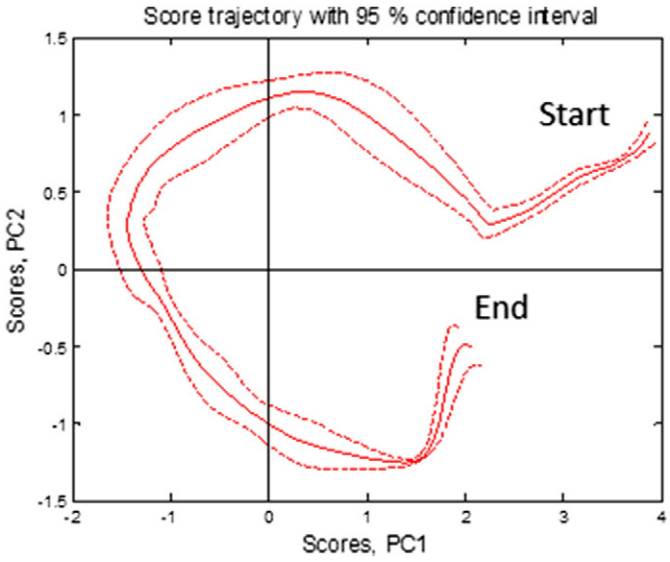

Fig. 5. The estimated trajectory with a 95% confidence interval.

contribute with positive values along PC1 (87% explained variance). Although the second PC explains only 12% of the variance, this is still systematic variation, thus the optimal model rank is two.

One way to model the batch data is to make use of a so-called maturity index where a response variable is constructed from the sample number for the various batches. Each sample is given a maturity index based on the length of individual batches and used as the response variable. However, this approach for modeling historical and predicting new batches does not solve the problem of non-linear behavior or differences in the starting point in relative time. As an illustration of this, a PLS regression model with a maturity index was calculated. A subset of the samples was selected so that they represent the same start and end point in chemical sense, as seen in the score plot (Fig. 3). Since the number of samples was not the same for all batches (from 147 to 158 observations), a relative maturity index from 1 to 100 was calculated as the average step based on the number of observations for each batch. Nevertheless, because of the highly non-linear behavior in the score space, this approach gave a poor explained variance for maturity as the response variable.

The data were then modeled with the new relative time procedure. Fig. 4 shows how the algorithm finds the grids as the basis for the estimation of the batch trajectory. The score space has been split in sections for PC1 and PC2, giving grids as depicted as rectangles. All samples found inside the individual grid elements serve as the basis for estimation of the mean values and the standard deviation. Interpolation is performed with the mean values of the scores inside the grids to produce a continuous trajectory with the desired resolution, e.g 100 points along the trajectory. Depending on the application, a linear or non-linear interpolation must be defined, Fig. 5. Note that the reaction is highly non-linear in terms of sample number versus relative time as e.g. the congested region in the left part of the plot shows. In fact, it seems that the reaction for a short period of time is actually reversed for batch 3 (in green).

Once the batch trajectory has been determined, all observations for the historical data can be represented in relative time irrespectively of the number of observations, thus avoiding the need to regulate sample frequency.

Fig. 6 shows the variable Temperature A in relative time as estimated from the score trajectory in the 2-dimensional model. It shows that the evolution of the batches now corresponds to the chemical time similar to the 2D score plot when plotted as a 1D control chart.

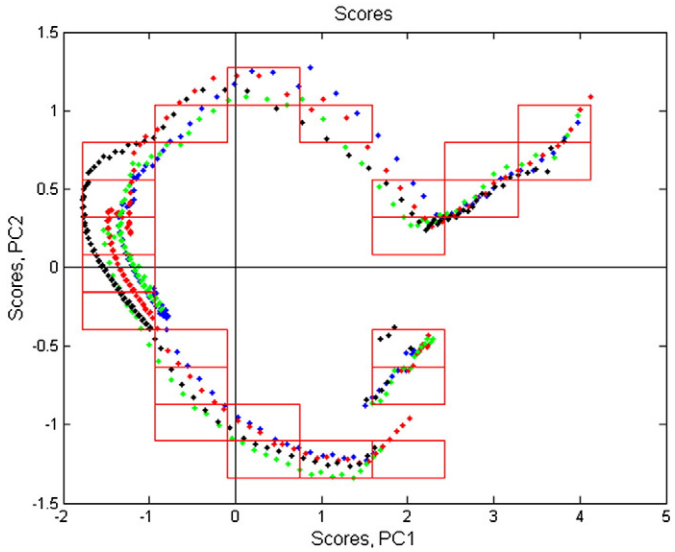

Fig. 4. Score plot for PC1 vs. PC2 with the grids depicted.

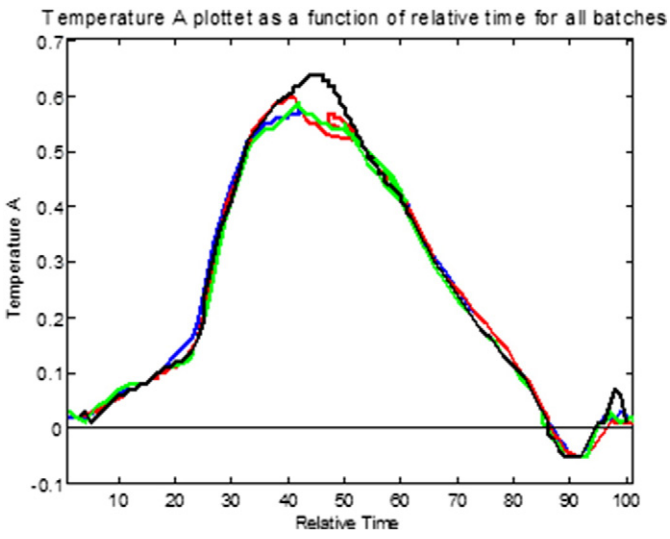

Fig. 6. The variable Temperature A in relative time.

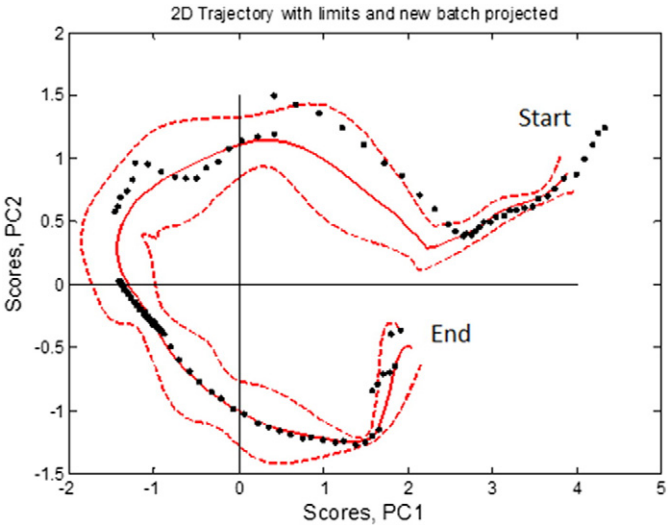

Fig. 7. Projection of a new batch on the chemical reaction batch trajectory model.

Projection of new batches is now straightforward in that new scores are estimated from the loadings and then projected onto the established batch trajectory. Fig. 7 shows how a new batch evolves over time (from right to left). Even if the batch starts outside the trajectory starting point because the chemical state is different, i.e. this batch has not progressed to the common starting point, this poses no problem in terms of visualizing in relative time. This is also the case in one dimension, the only impact is that the new observations may start in negative relative time. This is further discussed below for the fluid bed dryer data of example 2.

3.2. Example 2: Near-infrared (NIR) spectra of a fluid bed dryer operation

NIR spectra were collected during the fluid bed drying of a granular intermediate in a pharmaceutical process. The spectra consist of a total of 1093 wavelengths in the range 9090–4484 cm^{-1} (1100–2230 nm) at 4 cm^{-1} resolution. Due to the noisy nature of Fourier transform (FT) NIR (particularly at such high resolution), the spectra were smoothed with an 11-point moving average before applying a standard normal variate (SNV) to reduce the systematic baseline effects in the spectra. The spectra after SNV are shown in Fig. 8. A PCA model was calculated, showing that the two first PCs accounted for 97 and 1% of the variance, respectively. The validation variance indicated two PCs to be optimal which was confirmed by visual inspection of the 2D score plot.

Fig. 9 shows the historical batches with the grid elements found, and Fig. 10 shows the estimated trajectory with a 95% confidence limit and a new batch projected onto the model (black dots). When the batches are plotted as "sample number" in a 1-dimensional score plot (Fig 11) the initial interpretation is that

1. The historical batches are quite different.
2. The new batch is different from the historical batches.

However, the 2-dimensional score plots in Figs. 9 and 10 reveal that the batches do follow a common trajectory in relative time. Accordingly a line plot of the scores vs sample number gives an incorrect representation of how the batches evolve as the starting point for monitoring of the new batch cannot be assumed to represent the common starting point t_0.

The 2-dimensional score plot in Fig. 9 shows that the batches start in different positions because the starting material has various levels of moisture. This means that the common starting point is found where all batches follow the same trajectory, and the end point granulate was found by confirmation against a loss on drying reference analysis. Thus, the process was stopped when the moisture content presumably was at the correct level. Fig. 9 indicates that some batches were "overdried" as the end points in the relative time differ to some extent.

The scores for the new batch at the start of the process (lower left in Fig. 10) hardly change for the first 23 samples, but this has no impact on the progress in relative time. On the other hand, plotting these scores as sample number in Fig. 11 gives the impression that the new batch right from the start is out of control. The same is the case for the third historical batch (in green),

Fig. 10 shows the new batch starts outside of the common trajectory due to higher moisture than the common starting point for the historical batches. This situation can already be seen in Fig. 9, where some batches have their first observations in the lower left corner. However, as the granulate is dried the new batch follows the trajectory. Also notice the gap from one observation to the next as the batch enters the common trajectory. The reason for this may be heterogeneity of the granulate. Nevertheless, this does not pose any problem in monitoring the progress of the process.

4. Discussion

The results from the two examples presented above show that the new approach is able to model the batch progression in relative time,

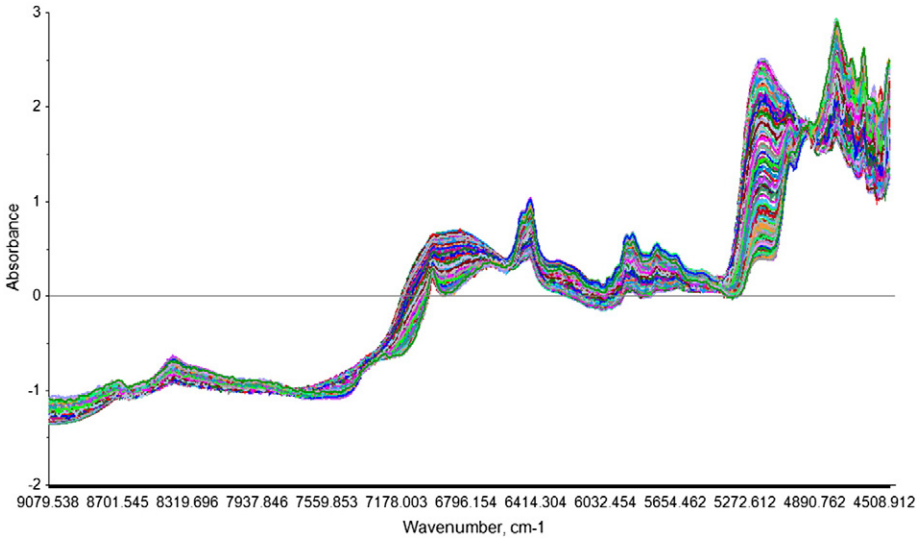

Fig. 8. NIR spectra after SNV.

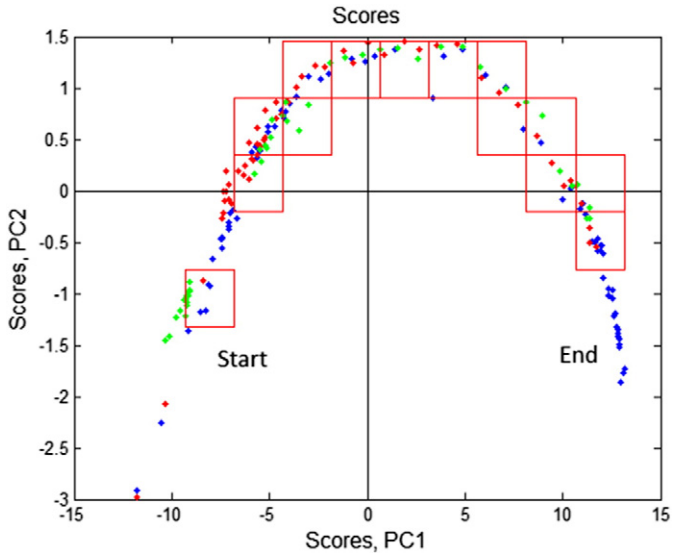

Fig. 9. Score plot of PC1 vs. PC2 for three historical batches with the grid elements shown.

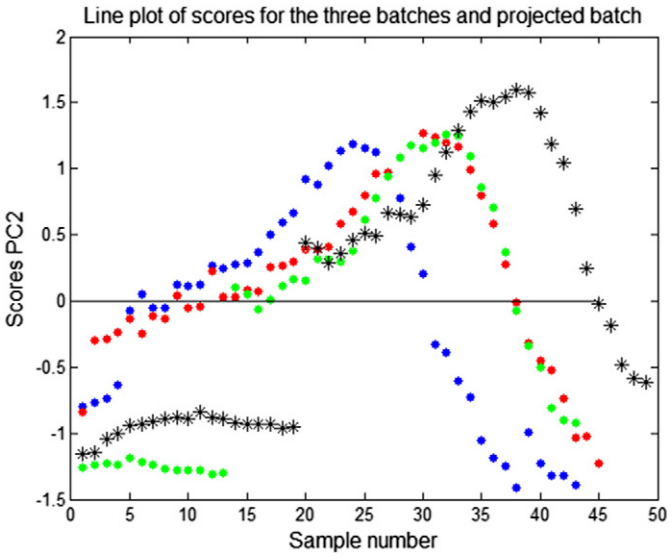

Fig. 11. Line plot of scores for PC2, historical batches as dots and the new batch as stars.

that better represents the batch progress compared to existing modeling approaches. The cross-validation over batches gives a conservative estimate of the explained validation variance as a way to decide on the optimal number of PCs. Figs. 7 and 10 depict the trajectory in 2D score plots; however, confidence intervals are also estimated for scores for individual PCs. If interpretation of the validation and visual assessment of the model concludes that a 1-dimensional model is optimal, the monitoring can still be performed in relative time.

The underlying assumption of the proposed approach is that the batch trajectories can be captured by a PCA model. From an assessment of the literature and from experience, this is true for most batch applications (primarily due to the highly correlated nature of consecutive points in batch models) as the data reflect the transition in the batch. If no feature space model can be found, the batch transition is not reflected in the data and it is doubtful that any of the existing batch modeling approaches would be successful.

In batch processes, there is typically a strong gradient from start to end. It can be imagined that the process is reversing, i.e. that the batch is revisiting an earlier feature state. As demonstrated in example 1, moderate reversing is handled by the relative time method. For more extreme reversing situations, adaptations of the approach might be required but that is outside the scope of this work.

In this paper, a static grid was used in the feature space to capture the systematic evolution of the batches. The static grid provides a robust and true representation of the presented examples. However, in more complex situations where e.g. heteroscedastic noise is present, a more flexible grid-search approach may be required.

Some batch processes, e.g. fermentation, are characterized by various phases transitions. In these situations, the correlation between the variables and the variance between the batches may not be constant for the whole duration.

Hierarchical models may prove useful in this case, where underlying classification models can be used to detect phase changes in an unsupervised manner and define a modeling and monitoring strategy based on joining numerous batch models together. In this situation, based on the location of the new time point in feature grid space, quantitative (or qualitative) models can be run in parallel where local refined models can be applied to assess the process, depending on the exact chemical/biological state of the material.

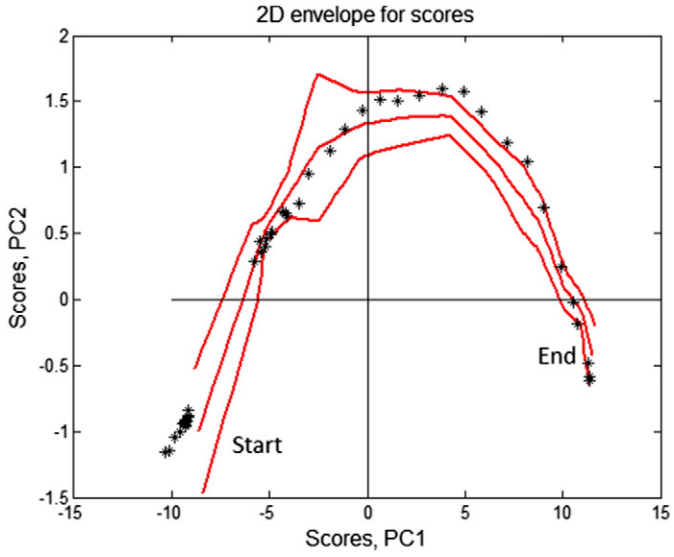

Fig. 10. Score plot of PC1 vs. PC2 with estimated trajectory from samples inside the grid elements, 95% confidence limit and a new batch projected.

5. Conclusions

The novel approach to batch modeling proposed in this paper models historical as well as monitors new batches directly in relative time. The common start and end points for the batches are found by a grid-search method. This eliminates the need for subjectively finding the samples that describe the trajectory in a common way, and thus, there is no need for time warping which may not be the correct remedy to handle various batch lengths and varying progression of the process in the first place.

Confidence intervals are estimated for the 2D and 1D score trajectories and can be extended to models with three (or higher) principal components (PCs). Monitoring of new batches is furthermore independent of the sampling rate allowing for changes in the sampling frequency between the NOC data set and the new batches.

The dynamic distance relative to the trajectory is estimated in a similar way as for Hotelling's T^2 statistic for PCA. This also extends to any model dimension. Dynamic Q-residual or F-residual critical limits are estimated along the trajectory for added diagnostic capability.

The individual variables can be presented as a line plot in relative time facilitating real-time monitoring for end users with no knowledge about multivariate methods. Overall, this approach represents an objective way of modeling complex systems such as biological fermentations where starting material attributes are highly variable by nature. The application of processing variables to such materials results in different manners that process can proceed. Eliminating time from the analysis (in a modeling sense) allows chemical/biological probing of the process in the state that the material exists in the process, without implying any need to fit the model to a certain time scale in terms of a given sample number.

References

[1] S. Wold, P. Geladi, K. Esbensen, J. Ohman, J. Chemom. 1 (1987) 41–56.
[2] P. Nomikos, J.F. MacGregor, AICHE J. 40 (1994) 1361–1375.
[3] J. Camacho, J. Pico, A. Ferrer, Anal. Chim. Acta 642 (2009) 59–68.
[4] X. Meng, A.J. Morris, E.B. Martin, J. Chemom. 17 (2003) 65–85.
[5] A. Kassidas, J. MacGregor, P. Taylor, AICHE J. 44 (1998) 864–875.
[6] C. Ündey, S. Ertunç, A. Çinar, Ind. Eng. Chem. Res. 42 (2003) 4645–4658.
[7] P.R.C. Nelson, P.A. Taylor, J.F. MacGregor, Chemom. Intell. Lab. Syst. 80 (2006) 1–12.
[8] J. Camacho, J. Pico, A. Ferrer, J. Chemom. 22 (2008) 299–308.
[9] A. Bogomolov, Chemom. Intell. Lab. Syst. 108 (2011) 49–63.
[10] H. Martens, M. Martens, Multivar. Anal. Qual. Wiley & Sons, 2001.